ARTÍCULOS

HOW WE DRAW TEXTS: A REVIEW OF APPROACHES TO TEXT VISUALIZATION AND EXPLORATION

Jaume Nualart-Vilaplana, Mario Pérez-Montoro y Mitchell Whitelaw

Nota: Este artículo puede leerse traducido al español en:
http://www.elprofesionaldelainformacion.com/contenidos/2014/may/02_esp.pdf

Jaume Nualart-Vilaplana is a PhD candidate in the *Faculty of Arts and Design*, *University of Canberra* (Australia), research engineer at *Nicta* (Australia), and a PhD candidate in the *Faculty of Information Science*, *University of Barcelona*. MAS and MSc (Licenciatura) at *Autonomous University of Barcelona*
http://orcid.org/0000-0003-4954-5303

Machine Learning Research Group at NICTA, Canberra Research Laboratory
Tower A, 7 London Circuit, Canberra City ACT 2601, Canberra, Australia
jaume.nualart@canberra.edu.au

Mario Pérez-Montoro holds a PhD in Philosophy and Education from the *University of Barcelona* and a Master in Information Management and Systems from the *Polytechnic University of Catalonia*. He studied at the *Istituto di Discipline della Comunicazione* at the *Università di Bologna* (Italy) and has been a visiting scholar at the *Center for the Study of Language and Information* (*CSLI*) at *Stanford University* (California, USA) and at the *School of Information at UC Berkeley* (California, USA). He is a professor in the *Department of Information Science* at the *University of Barcelona*. His work has focused on information architecture and visualization. He is author of the book *Arquitectura de la información en entornos web* (*Trea*, 2010).
http://orcid.org/0000-0003-2426-8119

Facultat de Biblioteconomia i Documentació, Universitat de Barcelona
Melcior de Palau, 140. 08014 Barcelona, España
perez-montoro@ub.edu

Mitchell Whitelaw is an academic, writer and practitioner with interests in new media art and culture, especially generative systems and data-aesthetics. His work has appeared in journals including *Leonardo*, *Digital creativity*, *Fibreculture*, and *Senses and society*. In 2004 his work on a-life art was published in the book *Metacreation: art and artificial life* (*MIT Press*, 2004). His current work spans generative art and design, digital materiality, and data visualisation. He is currently an associate professor in the *Faculty of Arts and Design* at the *University of Canberra*, where he leads the *Master of Digital Design*. He blogs at *The Teeming Void*.
http://orcid.org/0000-0001-9013-9732

Faculty of Arts and Design, University of Canberra
Bldg, Floor & Room: 9, C12. ACT 2617, Canberra, Australia
mitchell.whitelaw@canberra.edu.au

Abstract

This paper presents a review of approaches to text visualization and exploration. Text visualization and exploration, we argue, constitute a subfield of data visualization, and are fuelled by the advances being made in text analysis research and by the growing amount of accessible data in text format. We propose an original classification for a total of 49 cases based on the visual features of the approaches adopted, identified using an inductive process of analysis. We group the cases (published between 1994 and 2013) in two categories: single-text visualizations and text-collection visualizations, both of which can be explored and compared online.

Keywords

Review, Text visualization, Data visualization, Data exploration, Data display, Information visualization, Text analysis.

Título: Cómo dibujamos textos. Revisión de propuestas de visualización y exploración textual

Article received on 19-01-2014
Approved on 09-03-2014

Resumen

En este trabajo se presenta una revisión de estrategias para la visualización y exploración de textos. Se argumenta que la visualización y exploración de textos constituye un subcampo de la visualización de datos que se nutre de los avances en el análisis de textos y de la creciente cantidad de datos accesibles en formato texto. Proponemos una clasificación original para un total de cuarenta y nueve casos revisados. La clasificación está basada en las características visuales de cada caso, identificadas mediante un proceso inductivo de análisis. Agrupamos los casos (publicados entre 1994 y 2013) en dos categorías: las visualizaciones de texto individuales y la visualizaciones de colecciones de textos. Los casos revisados pueden ser explorados y comparados en línea.

Palabras clave

Visualización de texto, Visualización de datos, Exploración de datos, Visualización de información, Análisis de textos.

Nualart-Vilaplana, Jaume; Pérez-Montoro, Mario; Whitelaw, Mitchell (2014). "How we draw texts: a review of approaches to text visualization and exploration". *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 221-235.

http://dx.doi.org/10.3145/epi.2014.may.02

1. Introduction

The aim of this review is to propose a classification of text visualization and exploration tools, while describing the broader context in which they operate. To do so, we list, classify and discuss the most important contributions made in the field of text visualization and exploration between 1994 and 2013. This field is undergoing rapid growth –fuelled by open data initiatives and web scraping– and has become highly diversified, developing in parallel in a range of disciplines. Some of the most important visualization methods invented between 1765 and 1999 were the timeline, bar chart, pie chart, flow map, Venn diagram, histogram, Gantt chart, flowchart, tag cloud, social networks, boxplot, star plot, treemap, headmap, and sparkline. Figure 1 presents a word cloud (using *Wordle*) of the professions practiced by their respective inventors. Given this diversity, our search for cases has been conducted in many different contexts and has involved the examination of many different sources, ranging from the sciences to the humanities, from academic journals to blog sites, from universities to freelance studios, and from open data institutions to open data communities. Clearly this proliferation of disciplines has meant the adoption of a variety of different philosophies and points of view.

This review aims to help those that work with data, and especially with texts (but by no means limited to academics), to use visualization techniques that can identify patterns or behaviours present in the textual reality. Moreover, these techniques can help users improve –in terms of both the speed and the clarity of the process– the way in which they visualize and discover the facts that lie within the data.

Drawing a clear conceptual line between approaches to text visualization and exploration is no straightforward task, but here we have opted to review cases dedicated to both processes, be they described separately or together. Note that on occasions, for the sake of simplicity, we use the term text visualization in reference to both approaches.

The two types of text visualization considered here are:

1) Single-text representation, that is, ways of extracting meaning from texts based on writing style, document structure and language register as opposed to pure statistics. Our interest lies in representing the meaning and salient features of texts because their convenient visualization can speed up and/or improve our ability to select texts and manage the time required to tackle them. The research output of fields such as natural language processing, linguistic computing and machine learning provides techniques for producing high quality data representing complex texts. It is our belief that by combining these techniques with a suitable text visualization method we can improve the way in which we examine and understand texts.

2) Representation and exploration of collections of texts. Exploring and selecting individual texts and navigating and analyzing collections of texts are daily tasks for many of those who work with computers and datasets, and there is clearly plenty of room for new ideas and tools to facilitate their work. Information retrieval is a critical factor in an environment characterized by an excess of information (**Baeza-Yates** *et al.*, 1999). When a user conducts a search, the information retrieval systems normally respond with a list of results. More often than not, the presentation of these results plays an important role in satisfying the user's information needs, so a poor or inad-

Figure 1. Word cloud of the professions practiced by inventors of visualization methods

Table 1. Leading universities and their data visualization departments

Institution	Rank in 2012	Department/Course	URL
Harvard University	1	*Broad Institute of Harvard and MIT*	*http://www.broadinstitute.org/vis*
Massachusetts Institute of Technology	2	*Broad Institute of Harvard and MIT*	*http://www.broadinstitute.org/vis*
University of Cambridge	3	--	--
Stanford University	4	*Stanford Vis Group*	*http://vis.stanford.edu*
University of California, Berkeley	5	*VisualizationLab*	*http://vis.berkeley.edu*
University of Oxford	6	*Visual Informatics Lab at Oxford*	*http://oxvii.wordpress.com*
Princeton University	7	*PrincetonVisLab*	*http://www.princeton.edu/researchcomputing/vis-lab*
University of Tokyo	8	--	--
University of California, Los Angeles	9	*IDRE GIS and visualization*	*https://idre.ucla.edu/visualization*
Yale University	10	--	--

Table 2. Conferences dedicated primarily to data visualization ordered by number of participants (**Stefaner**, 2013)

Conference	Location	Topic	No. participants	URL
Nicar 2013	USA	Data journalism	149	*http://ire.org/conferences/nicar-2013*
Dd4d 2009	France	Information visualization	52	*http://www.dd4d.net*
FutureEverything 2013	UK	Technology/society/art	52	*http://futureeverything.org*
Resonate 2013	UK	Creative code	44	*http://www.thisisresonate.co.uk/resonate-13*
Graphical web 2012	Switzerland	Open web/datavis	38	*http://www.graphicalweb.org/2012*
IeeeVis - VisWeek 2012	USA	Information visualization	-	*http://ieeevis.org*
EuroVis 2013	Germany	Computational aesthetics	-	*http://www.eurovis2013.de*
Siggraph 2013	USA	Computer graphics and interactive techniques	-	*http://s2013.siggraph.org*
OzViz 2012	Australia & NZ	Workshops for visualisation practitioners, academics and researchers	-	*http://www.ozviz2012.org*

equate presentation can thwart the user (**Baeza-Yates** *et al.*, 2011). Typically, information retrieval systems present the results of a query in a flat, one-dimensional list. Such lists tend to be opaque in terms of the order they give to the information, i.e., the users are unaware as to why the list is presented in a particular order. To refine their search, users have to interact again, normally by filtering the first output of results. It is our belief that new techniques for representing collections of texts –including search results– can help improve navigation, exploration and retrieval.

As we show below, data visualization can today be considered a consolidated academic field (**Strecker**; *IDRC*, 2012). Thus:

- Seven of the top 10 universities according to the *Times Higher Education ranking* (2012) have departments or research groups working in the field of data visualization. The discipline is incorporated in a wide variety of departments, ranging from computer science and statistics to linguistics and graphical design, and from chemistry and physics to genetics and history. Recently, data visualization has emerged as a distinct field, with specific departments dedicated to its study and master's programs being taught in the subject (table 1).
- Over the last five years a number of conferences have been dedicated primarily to data visualization. These are listed in table 2.
- A number of journals are now specifically dedicated to studies in data visualization, and important contributions can be found also in conference proceedings (table 3).

Finally, a number of leading websites –including *Infosthetics*, *Visualcomplexity* and *Visualizingdata.com*– play a key role in the dissemination of the subject.

1.1. Text visualization

Shneiderman (1996) classifies regular texts as one-dimensional data, that is, data organized in a sequential manner, running right-to-left (or left-to-right), line-by-line, top-to-bottom. Yet, a text can have multiple internal structures, a morphology made up of paragraphs, sentences and words.

Table 3. Main journals dedicated to data visualization

Name	Url
Parsons journal for information mapping	*http://pjim.newschool.edu/issues/index.php*
Journal of visualization	*http://springer.com/materials/mechanics/journal/12650*
Ieee Transactions on visualization and computer graphics (TVCG)	*http://www.computer.org/portal/web/tvcg*
Information visualization	*http://ivi.sagepub.com*
International journal of image processing and data visualization (Ijipdv)	*http://iartc.net/index.php/Visualization*
IEEE Vis (former *Visweek*)	*http://ieeevis.org*
EuroVis	*http://www.eurovis2013.de*
ACM CHI	*http://chi2013.acm.org*
EG CGF	*http://www.eg.org*
IVS	*http://www.graphicslink.co.uk/IV2013*

Depending on its information structure, a text may be ordered by chapters, parts, sections, subsections, etc. If a text is given in a specific format, such as html, then it may be organized into bodies, divs, paragraphs, etc. In these examples the text includes tree structures as well as a one-dimensional structure. Additionally, texts may have a subjective component and an abstract structure that is not readily analysed by a computer. All in all, these data types and structures constitute the specificities of a text.

The amount of data to which we have access grows on a daily basis. Most of these data are in text format, as **Fernanda Viégas** and **Martin Wattenberg** in an interview with **Jeff Heer** argue: "One of the things I think is really promising is visualizing text. That has been mostly ignored so far in terms of information visualization approaches, and yet a lot of the richest information we have is in text format" (**Heer**, 2010).

> "Seven of the top 10 universities have departments or research groups working in data visualization"

Data analysis defines the boundaries of data visualization, i.e., it provides the fine line between multiple truths and lies. In the case of text visualization, this role has been taken on by text analysis: in the main, via computational linguistics, natural language processing, machine learning and statistics. The advances made in text analysis at a whole range of levels have provided computers with text understanding, enabling them to modify a text, the so-called unstructured data (see next subsection "Text analysis").

There is some discussion as to whether text visualization might be considered a specific subfield of data visualization. Some authors tend to disagree: **Illinski** (2013) claims that text cannot be considered a data type; **Šilić** (2010) argues that "unstructured text is not suitable for visualization". Yet, as discussed above, most text visualizations transform the initial "unstructured" textual data into a reduced, structured dataset. This new dataset is no longer one-dimensional, but rather it constitutes a categorical or a network dataset and it can be represented with a wide range of tools that are not specific to text representation (**Hearst**, 2009; **Grobelnik**; **Mladenić**, 2002).

As we show in the cases we review here, most text visualizations do not represent raw data: that is, the text as it is. Rather what they do is transform the text into smaller chunks of data, normally extracting a representative part of that text. This process is one of data transformation and it occurs, for example, when a text is reduced to a list of words based on their frequency of appearance. In that case, the method chosen to represent the data will belong to a family of methods best suited to the data type. In this review we consider the most frequently employed strategies to represent single texts or collections of texts, paying special attention to strategies for representing textual data as it is, as a regular text, with all its complexities, irregularities and rich abstractions.

Text analysis is a key field for text visualization. Below, we present a brief commentary on this matter and its relationship with text visualization.

1.2. Text analysis

Text analysis, roughly synonymous with text mining (**Feldman**; **Sanger**, 2006), is an interdisciplinary field that includes information retrieval, data mining, machine learning, statistics, linguistics and natural language processing. According to **Marti Hearst** (2003), the goal of text mining is to discover "heretofore unknown information, something that no one yet knows and so could not have yet written down". Text mining is a subfield of data mining whose typical applications include the analysis or comparison of literary texts, the analysis of biological and genomic data sequences and, more recently, the identification of consumer behaviour patterns or the detection of the fraudulent use of credit cards. **Hearst** differentiates these applications from information extraction operations, such as the extraction of people's names, addresses or job skills. This latter task can be done with >80% accuracy, but the former, the full interpretation of natural language by a computer program, looks like it will not be possible for "a very long time" (**Hearst**, 2003).

To study text visualization and exploration it is important to examine the literature dedicated to both data visualization and text analysis, given the significant interrelationships that exist. Thus, while the text analysis output may limit the possibilities of visual presentation and interaction with the text, there is strong empirical evidence indicating that people learn better with a combination of text and illustration (visualization) than with text alone (**Anglin** *et al.*, 2004; **Levie**; **Lentz**, 1982).

2. Review

In this section we propose a possible classification based on the visual features that characterize the approaches to textual visualization and exploration, as identified in 49 cases.

The methodology to collect the cases is a two-part process. First, a traditional literature search and review (including practical examples and visualisation studies); and second, a subset of these have been selected, based on a preliminary analysis of their features. The aim was to select cases that provided a representative overview of the range of work in the field.

The classification of the cases is the product of empirical observation following an inductive analysis. The classification is followed by an analysis of these cases.

There are alternatives to those used in this paper for the selection and categorization of primary source methodologies such as **Kitchenham** (2004) and **Benavides**; **Segura**; **Ruiz-Cortés** (2010).

2.1. Classification of approaches

The basic classification of text visualization approaches comprises two categories according to the type of data to which they are applied:

1) Textual documents: that is, representations of single texts, where text is understood as a sequence of words ordered according to the hierarchy: document > paragraphs

> sentences > other punctuation marks > words > syllables and phonemes or morphemes. Where a text is a book or another kind of structure, then, it may have more granularities, including: chapters > sections > sub sections > etc. We also include the metadata of the text and other attached texts, i.e., title, author(s), publisher, copyright notes, acknowledgement, dedication, preface, table of contents, forward, glossary, bibliography, index, etc.

2) Text collections: that is, a group of texts in which each item constitutes a clearly differentiable entity. Typically when speaking of collections of texts, we speak of texts that have elements in common, be it their register, length or structure. All the cases we review here are collections of the same text type. Heterogeneous collections of texts are also referenced in the literature (**Meeks**, 2011), especially in representative analyses of a field of knowledge, where the aim is to include the greatest possible variety of expressions and vocabulary. In such cases the dataset can be said to be heterogeneous in term of its structure and register.

To these two data types, we then add several subjective subdivisions to each category according to the visual features used to represent the textual features. The aim here is to be able to describe and explain the cases under review, as well as to identify the key features of the text visualization approaches.

Single texts
- Whole <-> Part
- Sequential <-> Non sequential
- Discourse structure <-> Syntactic structure
- Search
- Time

Text collections
- Items <-> Aggregations
- Landscape
- Search
- Time

2.1.1. Single texts

In the specific instance of single texts, we classify the cases according to the part of the text that is represented, whether the approach follows the same sequence as that of the text, and the text structure employed in each case.

Whole or part?

In some instances, one part of the text is considered the essence of the text and is used in the visualization process rather than the whole text. Yet, there are processes that use the whole text, at least implicitly. Examples include:

- chapters of a book but not the whole text.
- representation of all the sentences of the text as coloured lines.
- verbs of a text, providing an impression of the style of the text.
- characters of a novel and their appearance within the text.
- places or dates present in the text.
- etc.

The cases in which the whole text is explicitly represented are, for obvious reasons, cases involving relatively short texts, e.g., song lyrics, speeches, poems, etc.

In some instances, such as when using *Radial word connections* (see, case 1 below) only certain words from the text are represented; yet, we classify this case as a whole text representation because the whole novel, chapter by chapter, is implicitly represented in the circle.

In those instances in which the whole text is represented (even implicitly) as one central element in the visualization, we classify it as being a whole-text visualization.

Does the visualization follow the same sequence as that of the text?

If the visualization follows the same sequence, or order, as that of the text, then the case is considered sequential; if not, then it is considered non-sequential. For example, a typical case that does not follow the same sequence as that of the original text would be a word cloud (see figure 1).

> “Most text visualizations transform the initial ‘unstructured’ textual data into a reduced structured dataset”

Does the visualization use elements from discourse structure or from syntactic structure?

A text may present one of two kinds of structure that we consider useful for our research. One is so-called discourse structure. Depending on the nature of the text, the discourse structure can be completely subjective to the author’s point of view –as in literature–, or restricted to a given structure –as in legal and scientific texts. In linguistics, discourse is a broad concept, but here we use it to refer to the parts of a text and the outline of a document: parts, chapters, sections, subsections, etc. The discourse structure is widely used when visualizing texts because it is a relatively straightforward way to represent the text sequence.

The second structure is the text’s syntactic structure, referred to text structure in sentences, phrases and word classes —including verbs and nouns. This is an objective structure and is dependent on the rules of linguistics. In text visualizations, the elements comprising this structure, such as sentences, are very common.

2.1.2. Text collections

In the specific instance of text collections we classify the cases according to pure items or aggregations, i.e., as pure data or data landscapes. Thus we determine whether the items making up the collection can be differentiated or represented as aggregations. The specific questions we address are: How is each item in the collection graphically represented? Is each text represented as a graphical entity, i.e., as a point, a word or short sentence? Can the items in the visualization be counted, i.e., are they visually differentiated?

There are cases in which each item is not represented by a graphically distinct entity, but rather, for example, as a coloured block. Alternatively, the items are accumulated and shown as frequency distributions. When the items of the collection are not graphically distinct (visually countable)

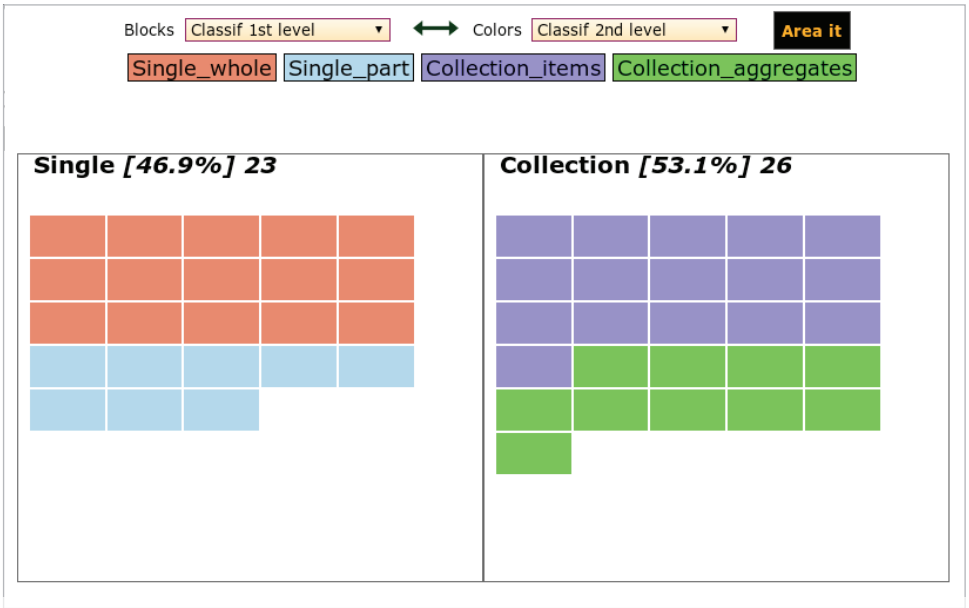

Figure 2. The 49 reviewed cases visualized with the *Area* software (screen shot).

then we speak in terms of the visualization of an aggregation rather than that of an item.

Pure data or data and landscape?

Are the items of the collection accompanied by any graphical content? Is another dataset, apart from that emanating from the text, also being represented? Some cases present the items embedded in a graphical environment, such as a map. This context might be an actual geographical map, a metaphor, or, for example, a conceptual landscape composed of words that form a second layer complementing that of the data collection, in which every distance plays a role: item-item (similarity between documents), word-item (importance of a word in a document), word-word (similarity between words in the collection).

Scales and axes are not considered as landscapes, nor are the elements of the interface in which the representation is embedded. This data layer, if not considered as the main dataset, would reduce substantially **Tufte**'s data-ink ratio (**Tufte**; **Graves-Morris**, 1983) compared to the ratio of a pure data representation.

2.1.3. Both single texts and text collections

Properties that are equally applicable to single-text and text-collection visualizations include time, search results and dataset size.

Does time play a role?

Do the texts change over time? One set of visualization approaches highlights the changes undergone by a dataset over time. The most common approaches of this kind have been developed in computer science to represent code evolutions or in *Wikipedia* to indicate various aspects of article revisions.

This category also includes visualizations in which the dataset itself changes over time; for example, the visualization of the latest news will see the dataset grow over time.

Does the visualization result from a search query?

Visualizations of the output of information system retrieval is a well-defined kind of visualization characterized by the changing number of represented items depending on the number of search results obtained. This is a growing visualization subfield related to the disciplines of information systems and information retrieval (**Mann**, 2002; **Hearst**, 2009).

Validity for small or large datasets

It is rare that a visualization tool is independent of the size of the dataset that is to be represented. Here, in those cases in which the tool has been clearly designed for a specific dataset size, the reader will be given the corresponding explanation.

2.2. Analysis of visualization approaches

We review a total of 49 cases applying the classification outlined above. In an attempt to incorporate the most crucial aspects of text visualization, our review concentrates on the specific ideas underpinning the text visualization, rather than the dataset and the contexts of each case.

Sixteen fields have been collected for each case: name, short name, author(s), year of publication, URL for further information, original dataset, discipline related to the work, description of the visualization method, description of the case, screen shot, thumbnail, classification (single or collection), classification (single-whole, single-part, collection-items, collection-aggregations), classification (time), classification (search), classification (dataset small, dataset large, N/A).

The cases are grouped into two sections and four subsections:

Single-text visualizations (23 cases)

– Whole-text visualizations (15 cases)
– Partial-text visualizations (8 cases)

Text collection visualizations (26 cases)

– Collection of items (16 cases)
– Collection of aggregations (10 cases)

For each subsection the cases are sorted by year of publication (descendant). To assist the reader, the collection of all reviewed cases can be viewed using the visualization and exploration software (also included in the review) known as *AREA* (**Nualart**, 2013).

2.2.1 Single-text visualization

We present single texts grouped as whole-text visualizations, partial-text visualizations and other subcategories.

The latter includes sequential and non-sequential visualizations, discourse-structures and syntactic-structures visualizations, search results and datasets dependent on time visualizations. Each subsection adheres to the following structure: list of cases, description of the group and discussion.

a) Whole-text visualizations

1) Literature. *Novel views: Les misérables*, *Radial word connections* by Jeff Clark (2013)

2) Literature. *Novel views: Les misérables*, *Character mentions* by Jeff Clark (2013)

3) Literature. *Poem viewer* by Katharine Coles *et al.* (2013)

4) Politics. *State of the Union 2011*, *Sentence bar diagrams* by Jeff Clark (2011)

5) Literature. *Visualizing lexical novelty in literature* by Matthew Hurst (2011)

6) Science/papers. *On the origin of species: The preservation of favoured traces* by Ben Fry (2009)

7) Science/papers. *Texty* by Jaume Nualart (2008)

8) Religion. *Bible cross-references* by Chris Harrison (2008)

9) Literature. *Literature fingerprint* by Daniel A. Keim and Daniela Oelke (2007)

10) *Wikipedia. History flow* by Fernanda Viégas and Martin Wattenberg (2003)

11) Literature. *Colour-coded chronological sequencing* by Joel Deshaye and Peter Stoicheff (2003)

12) Literature. *2-D display of time in the novel* by Joel Deshaye (2003)

13) Literature. *3-D display of time in the novel* by Joel Deshaye (2003)

14) Any. *Wattenberg's arc diagram* by Martin Wattenberg (2002)

15) Health. *TileBars* by Marti A. Hearst (1995)

Description

- Number of cases: We identify 15 cases that can be categorized as whole-text visualizations.
- Years: The cases were published over an 18-year period from 1995 to 2013.
- Authors: All the authors work in academic fields. The most prolific authors in this category are Jeff Clark and Joel Deshaye (with three cases each), followed by Martin Wattenberg (with two cases).
- Datasets: Most of the text corpora in this category are taken from literature (eight cases). Most authors draw on novels, especially well-known texts such as the classics, to demonstrate new visualization approaches.
- Methods: All the cases except case 14 (*arc diagram*) use colour as part of the visualization method. Five cases use methods that are bar chart derivatives (cases 4, 5, 6, 9 and 11). Three cases use curves connecting parts of the texts: two arcs and one radial diagrams (cases 1, 8 and 14).

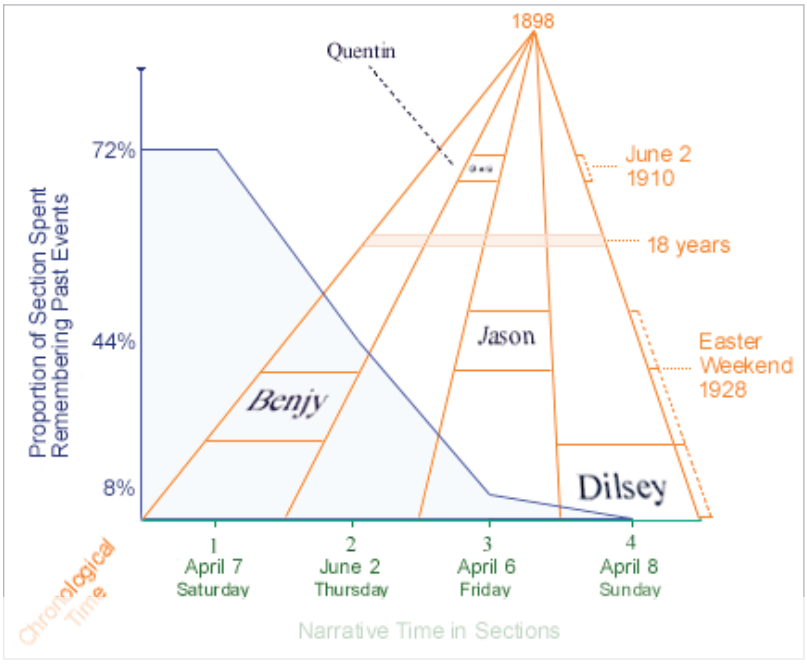

Figure 3. (Case 13) *3-D display* of timeof William Faulkner's novel *The Sound and the Fury*, by Joel Deshaye and Peter Stoicheff (2003)

Discussion

A common method cannot be identified for these whole-text visualizations. Yet, as expected, they all present an axis representing the whole text. In 13 of the 15 cases, the text line is represented by a horizontal or vertical line. The two exceptions use a circle –the case of *Radial word connections* (case 1)– and an iconification of a text on the page –the case of *Texty* (case 7).

Since whole-text visualizations always include an abstraction of the text, referred to as its text line, a question arises: which part of the text is physically present in the whole-text visualization being reviewed? Interestingly, nine of the 15 visualizations do not show a single word (cases 4, 5, 6, 7, 8, 9, 10, 11 and 15). Four cases show a small number of words (cases 1, 2, 12 and 13) (figure 3), while only two cases show all the text (cases 3 and 14).

The most common approach is to show the occurrence of a certain feature –this might be a term, topic, cross-reference or character– within the text as a whole (all cases except 3, 12, 13 and 15). With the exception of Wattenberg's *arc diagrams* (case 14), these occurrences are represented using the same colour.

It is interesting to observe how very similar data are represented in very different ways depending on the case under review. For example, while Viégas and Wattenberg's *History flow* (case 10) and Fry's Favoured Traces (case 6) both present document-version histories by section, the former is spatialized and the latter animated. Similarities, however, are seen in the approaches adopted, for example, by *TileBars* (case 15) and *Texty* (case 7). Thus, both highlight words from the text within a rectangular figure that is representa-

tive of the whole text. Other cases use opposite or complementary techniques. Thus, Wattenberg's *Arc diagram* (case 14) shows repetitions while Hurst's novelty visualization (case 5) shows only new strings, and no repetitions.

Literature and other complex texts, such as political speeches (case 4) and the *Bible* (case 8), dominate the type of corpora used in this category (10 cases). This is perhaps surprising, as these texts tend to be complex, often presenting a high level of abstraction and little formal structure. Arguably, when opting to introduce or test a new approach, it would make more sense to work with simpler, more structured texts (such as scientific papers, patents, health diagnostics, etc.) that present greater regularity in terms of their vocabulary, text length, discourse structure and register. Given the inherent freedoms associated with literature, novelists are under no obligation to adhere to any pattern or rule that might help us give structure to the unstructured.

However, depending on how the text is treated and processed, the nature of the text is not always relevant. For example, Matthew Hurst (case 5) tracks the introduction of new terms in literary texts. Yet the tool can be applied to any other text type, its results being unrelated to the complexity of the text given the ubiquity of the method. Having said this, it would be interesting to apply the technique to scientific papers in which the style is much more clearly defined. Similar arguments can be applied to *Radial word connections* (case 1), *Sentence bar diagrams* (case 4) and *Literature fingerprints* (case 9).

b) Partial-text visualizations

16) Literature. *Novel views: Les misérables. Characteristic verbs* by Jeff Clark (2013)

17) Any. *Wordle* by Jonathan Feinberg (2009)

18) Books. *DocuBurst* by C. Collins, S. Carpendale and G. Penn (2009)

19) Literature. *Phrase nets* by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

20) *Google* data. *Word spectrum: Visualizing Google's bi-gram data* by Chris Harrison (2008)

21) *Google* data. *Word associations: Visualizing Google's bi-gram data* by Chris Harrison (2008)

22) Literature/songs. *Document arc diagrams* by Jeff Clark (2007)

23) Any book. *Gist icons* by P. DeCamp, A. Frid-Jimenez, J. Guiness, D. Roy (2005)

Description

- Number of cases: We identify eight cases that can be categorized as partial-text visualizations.
- Years: The cases were published over an eight-year period from 1995 to 2013.
- Authors and datasets: Two cases by Jeff Clark (cases 16 and 22) and one by the creative team of Wattenberg and Viégas in collaboration with van Ham (case 19) use literary texts. The two cases by Chris Harrison use large *bi-gram* datasets published by *Google*. One case is not dependent on the nature of the text: *Wordle* (case 17), the very popular "word cloud" method introduced by Feinberg. Finally, two interactive approaches involving large datasets are presented: *DocuBurst* (case 18) and *Gist icons* (case 23).
- Methods: In six of the eight cases (cases 16, 17, 18, 19, 22 and 23), the dataset is reduced to what is called a bag of words and only these words are present in the visualization. Cases 20 and 21 are representations of all bi-grams that pit two primary terms against each other.

Discussion

Partial-text visualization is a successful, popular way to draw a text, presumably because of the way in which a long text can be effectively represented using a small set of words. Simple statistical methods such as word frequency counts are readily interpretable. A list of variously sized words is a direct way of communicating with any user, from beginner to expert. Most of the partial-text approaches available online use statistical methods to extract the part from the whole.

It is our contention that extracting part of the corpora can be affected by the structure and complexity of the whole. In the visualizations under review, half present unstructured text corpora, but the criteria used in extracting the part from the whole are well defined and include lists of verbs (*Characteristic verbs*, case 16), words occurring in the text in an "X and Y" pattern (*DocuBurst*, case 18) and lists of words not included in a list of predefined empty words (*Google's bi-gram data*, case 21).

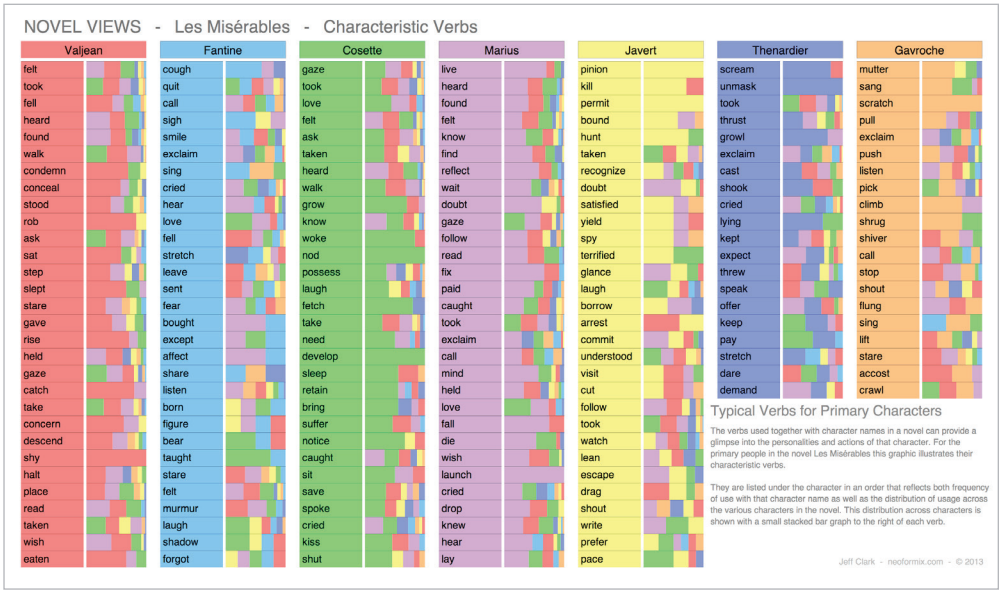

Figure 4. (Case 16) *Novel views: Les misérables. Characteristic verbs* by Jeff Clark (2013)

Clearly, extraction processes based on word or phrase functionality, as opposed to those that use statistical methods, are more closely affected by the nature of the text. Here, we focus on these cases because they are more interesting in terms of our research goals. They include the cases of *Novel views: Les misérables*. *Characteristic verbs* (case 16), which represents only verbs, *DocuBurst* (case 18) which uses the crowd-sourced lexical database *Wordnet* as a human-like backup, and *Phrase net* (case 19) and the two *Google bi-gram visualizations* (cases 20 and 21).

A common pattern detected in the partial-text visualizations reviewed is that once a part of the text has been extracted all except one (*Document arc diagrams*, case 22) discard any reference to the original text sequence in the visualization. See the following point for a more detailed discussion of this idea.

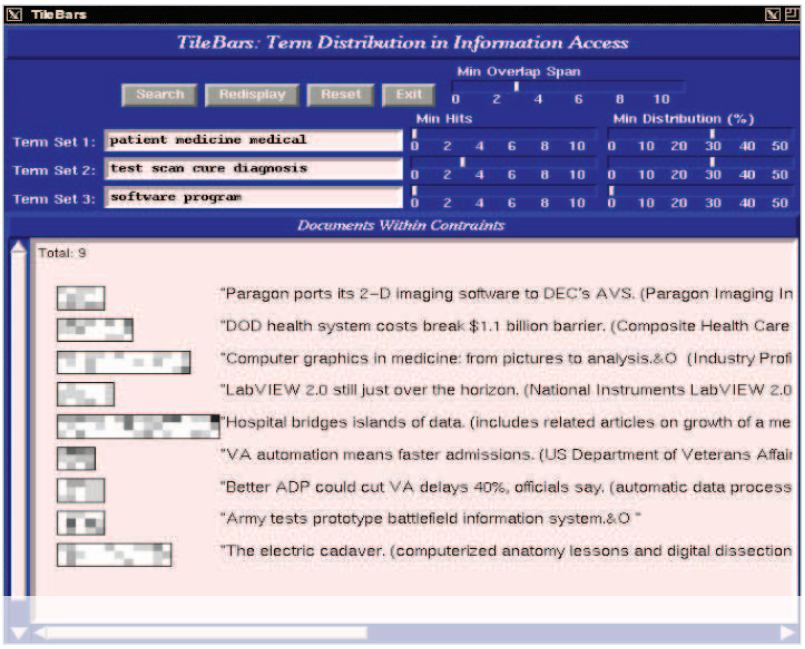

Figure 5. (Case 15) *TileBar* search on (patient medicine medical AND test scan cure diagnosis AND software program) with stricter distribution constaints.

c) Other subcategories

Here we include sequential and non-sequential visualizations, discourse and syntactic structures visualizations, search results and datasets dependent on time visualizations.

Sequential visualizations

Sixteen of the 23 single-text visualizations maintain a similar sequence to that of the original text. Seven of these visualize the sequence using a discourse structure (primarily chapters), while the remaining nine use syntactic elements to represent the sequence of the text (e.g., words).

Strikingly, only one partial-text visualization, Clark's *Document arc diagrams* (case 22) (figure 5), follows the original text sequence, whereas all the whole-text visualizations are sequential. It would thus appear that sequentiality is intrinsic to whole-text visualization. Whole-text visualizations do not literally represent every word of the text, but rather present a graphical metaphor of the whole: a text line. This text line may represent either a discourse structure or a syntactic structure of the text; but, whatever the case, graphically a line or area is used to represent the length of the text.

The sequentiality of the visualization means it can be read both backwards and forwards, as can the text. In the case of a long text, such as a book (nine of the 16 cases), the visualization can serve as a map or guide to the text.

Non-sequential visualizations

Five cases use non-sequential visualizations: three use word clouds (cases 17, 20 and 21), one a net of phrases (case 19) and one visualizes all the verbs in the text (case 16).

Discourse structures in the visualization

Cases: 1, 2, 5, 6, 8, 11, 12 and 13

The eight visualizations that follow the discourse structure of the text are sequential –no cases being found in which the discourse structure appeared out of sequence with regards to the text. This is perhaps unsurprising, as those cases in which the text is divided into chapters and each chapter represented as a separate entity were considered as text collection visualizations (e.g., *Sentence bar diagrams*, case 4). For this reason, all the cases in this section represent the parts of a text ordered and aligned (in a curve or line). Of the eight visualizations, five represent chapters or sections of a book, two represent complete volumes, while one (*Colour-coded chronological sequencing*, case 11) divides the text into paragraphs (alternatively with pictures and text). Indeed, case 11 is the only one we have identified that uses discourse structure elements that are more deeply embedded than chapters, sections, books and volumes. In all likelihood, more deeply embedded methods than these, such as, narrative topics, would require manual text line segmentation.

Syntactic structures in the visualization

Cases: 3, 16, 4, 7, 18, 9, 22 and 23.

The other eight sequential visualizations use intrinsic text elements, including groups of words (cases 7, 18, 22 and 23), verbs (case 16), sentences (cases 4 and 9) and a complete text analysis (case 3). Syntactic analysis requires either word-by-word parsing of the text (using a database of lexical or semantic word lists) or sentence and paragraph parsing. Syntactic-structure visualization is less dependent on the nature of the text in the sense that the methodology is unaffected by the complexity of the text. Typically, the software automatically extracts or marks the chosen syntactic elements.

Search-result visualizations

Cases: 15, 18 and 23

The three search-result visualizations were presented as web applications and were, therefore, interactive – the user being able to query the visualization system and obtain a

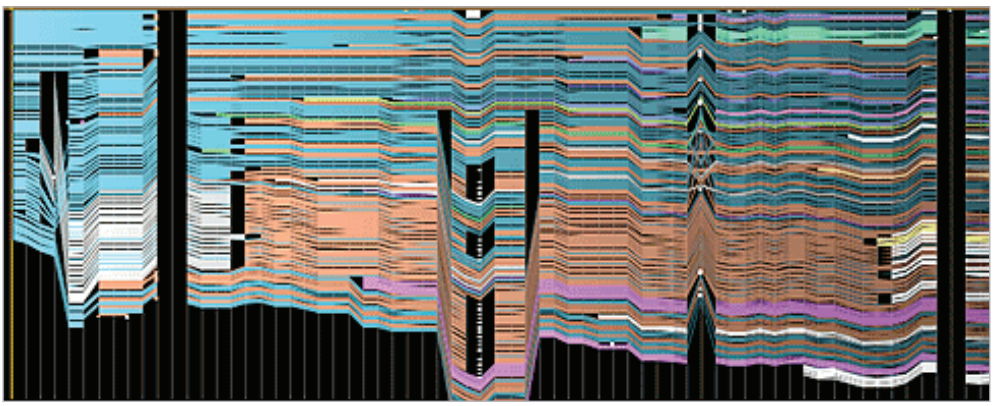

Figure 6. (Case 10) *History flow* by Fernanda Viégas and Martin Wattenberg researchers at *IBM's Visual Communication Lab* (2003)

unique representation for each search. The three cases, however, are no longer available online. *DocuBurst* (case 18) is a *Prefuse* application that can be downloaded (**Collins** *et al*., 2009). *Prefuse* is a set of software tools for creating rich interactive data visualizations.

TileBars is a classic case of visualization (cited 625 times by *Google Scholar*) designed by a leading expert in visualization and search engine interfaces, Marti Hearst. *DocuBurst* and *Gist icon* are interactive radial visualizations, the latter being one of the references and main influences on the development of *DocuBurst*, as explained in the *DocuBurst* paper cited.

> Partial-text visualization is a successful, popular way to draw a text, presumably because of the way in which a long text can be effectively represented using a small set of words

Search-result visualization approaches have not been widely implemented in information retrieval systems and most result outputs are one-dimensional lists of itemized texts (**Nualart**; **Pérez-Montoro**, 2013). The three cases reviewed here are each applied to large datasets and, starting with a search query, present an improved search output designed to help the user read and filter the results. All three are particularly concerned with distinguishing between similar items: *TileBars* searches *PubMed* (more than 20 million papers); *DocuBurst* uses the *WordNet* lexical database (155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs) to classify the visualized text; and, *Gist icons* use, among others, the complete dataset of approximately 7 million *USpto* patents and the *Enron* email dataset comprising 500,000 emails.

In the text collection category below, we present nine further search-result visualizations.

Time dependent datasets

Cases: 6 and 10.

We present two cases in which the visualization approaches can be used to understand or follow the evolution of a text over time. A dynamic text visualization demonstrates that data visualization may be the only way to solve certain tasks and that it is not just one more method of pure data advocacy. For example, it is extremely challenging to show how a *Wikipedia* entry evolves over time in line with the editors' participation (*History flow*, case 10) (figure 6). *History flow* provides a solution to this problem and sheds light on the complex collaborative process of *Wikipedia*.

In the second case (*Favoured traces*, case 6), an animated visualization demonstrates how Darwin's ideas evolved through successive editions of the *Origin of Species*. In Ben Fry's words: "The first English edition was approximately 150,000 words and the sixth is a much larger 190,000 words. In the changes are refinements and shifts in ideas —whether increasing the weight of a statement, adding details, or even a change in the idea itself."

2.2.2. Text collections

We present text collections grouped as pure item visualizations, aggregation visualizations and other subcategories. The latter includes data as a landscape layer and search result visualizations. Each subsection adheres to the following structure: list of cases, description of the group and discussion.

a) Item visualizations

24) Literature (Note: this converts a single text into a collection). *Novel views: Les misérables. Segment word clouds* by Jeff Clark (2013)

25) Literature. *Grimm's fairy tale network* by Jeff Clark (2013)

26) *Twitter*. *Spot* by Jeff Clark (2012)

27) Science. *Word storm* by Quim Castella and Charles Sutton (2012)

28) Literature. *Topic networks in Proust. Topology* by Elijah Meeks and Jeff Drouin (2011)

29) *Wikipedia*. *Notabilia* by D. Taraborelli, G. L. Ciampaglia and M. Stefaner (2010)

30) Media art. *X by Y* by Moritz Stefaner (2009)

31) Search engine. *Search clock* by Chris Harrison (2008)

32) Online media. *Digg rings* by Chris Harrison (2008)

33) Science. *Royal Society Archive* by Chris Harrison (2008)

34) *Wikipedia*. *WikiViz: Visualizing Wikipedia* by Chris Harrison (2007)

35) Visualization. *Area* by Jaume Nualart (2007)

36) *Chromograms* by M. Wattenberg, F.B. Viégas and K. Hollenbach (2004)

37) Search engines. *KartOO/Ujiko* by Laurent Baleydier and Nicholas Baleydier (2001)

38) Search engines. *Touchgraph* by TouchGraph, LLC. (2001)

39) Internet. *HotSauce* by Ramanathan V. Guha (1996)

Description

- Number of cases: We identify 16 cases that can be categorized as item visualizations.
- Years: The cases were published over a 17-year period from 1996 to 2013.
- Authors: The most prolific authors in this category are Chris Harrison (cases 13, 32, 33 and 34) and Jeff Clark (cases 24, 25 and 26), followed by Moritz Stefaner with two cases (29 and 30).
- Disciplines and datasets: Interestingly, nine cases are datasets taken from the Internet: *Wikipedia* (cases 29, 34 and 36), search engines (cases 31, 37 and 38), *Twitter* (case 26), online media (case 32), web pages (case 39). Only three cases use literary texts (cases 24, 25 and 28). Finally, two cases visualize scientific papers (cases 27 and 33), one case uses media art datasets (case 30) and one represents non-specific collections (case 35).

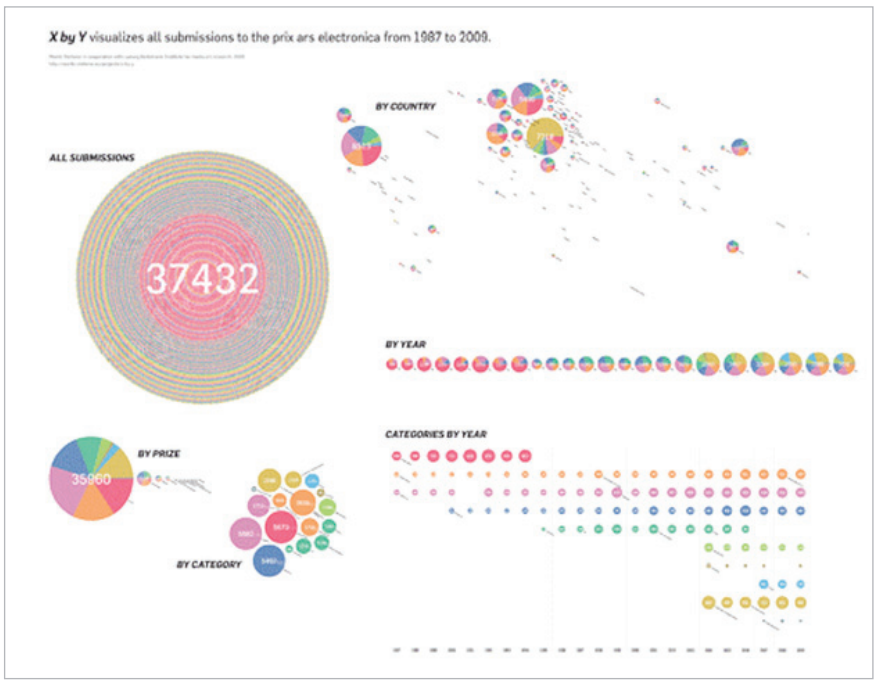

Figure 7. (Case 30) *X by Y* by Moritz Stefaner (2009)

Discussion

The main difference between single-text and text-collection visualizations lies in the nature of the text. In the case of the latter, most of the texts do not originate from literature and are accessible online. Yet, the nature of the text appears to be less important when the goal is the representation of the collection rather than of the text itself.

Item visualizations use methods that are independent of the nature of the items themselves. Once the text collections have been itemized, the dataset can be considered a general case of data visualization and not a pure case of text visualization. For this reason, in this category, the methods are generally well known and used in other fields of visualization. Thus, we find six network visualizations (cases 25, 28, 34, 37, 38 and 39), three timelines (cases 31, 32 and 33) and three cases that likewise use timelines but which also permit categorization-based groupings (cases 26, 30 and 35) (figure 7).

Finally, four cases are, we believe, quite specific to text visualization. Two are concerned with item comparison: *Segment word clouds* (case 24) and Word storm (case 27). *Segment word clouds* transforms a single text into a text collection. Specifically, it is used to represent the chapters of *Les misérables* as word cloud items, thus facilitating their comparison. It also uses colour to identify words as they acquire prominence in the text.

Word storm is a reinvention of word cloud, or more specifically a variation of *Wordle* (case 17) that allows word clouds to be compared. This is achieved by assigning a fixed position to each word. This simple idea makes it visually easy to compare word clouds while maintaining the usual word cloud features.

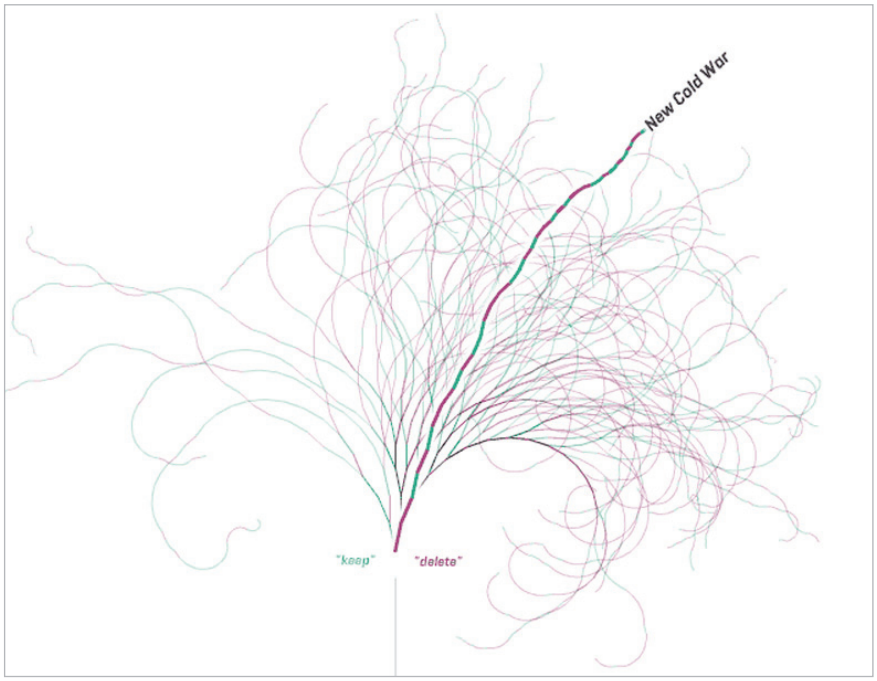

Figure 8. (Case 29) *Notabilia. 100 longest Article for deletion [AfD] discussions on Wikipedia* by Dario Taraborelli, Giovanni-Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization) (2010)

Figure 9. (Case 43) *Web seer* by Fernanda Viégas & Martin Wattenberg (2009)

To conclude, *Notabilia* (case 29) and *Chromograms* (case 36) are two highly original cases that deserve mention. The very specific design of *Notabilia* shows the evolution of "Article for deletion" discussions of Wikipedians (figure 8), discussions that are sometimes more like "flame wars" given the controversies that rage over the simple existence of certain definitions. *Notabilia* visualizes the evolution of the hundred longest discussions and their final outcomes. Moritz Stefaner's visualization constitutes an interactive bushtree, the branches of which are highlighted when moused over. The shape of the branches informs the reader about the nature of the discussion: cyclical, straight or never-ending.

Chromograms is also based on *Wikipedia* data, providing an analysis of the comments of editors for each edition of a *Wikipedia* entry. Visually it produces colour-coded stripes that in a small space rapidly inform the reader about the edit history of *Wikipedia* entries.

> It might prove more effective to apply visualization techniques to texts that have a more formal register and/or predefined outline and a well-defined vocabulary

b) Aggregation visualizations

40) Literature. *Grimm's fairy tale metrics* by Jeff Clark (2013)

41) Topic models. *Termite* by J. Chuang, C.D. Manning and J. Heer (2012)

42) *Wikipedia*. *Pediameter* by Müller-Birn, Benedix and Hantke (2011)

43) *Google* suggestions. Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

44) *Google* n-grams. *Web trigrams: visualizing Google's trigram data* by Chris Harrison (2008)

45) Political speech. *FeatureLens* by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman and C. Plaisant (2007)

46) Online news. *Newsmap* by Marcos Weskamp (2004)

47) Email conversation. *Themail* by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48) Search engine. *WebBook* by S.K. Card, G.G. Robertson and W. York (1996)

49) Any texts. *Dotplot applications* by Jonathan Helfman (1994)

Description

- Number of cases: We identify 10 cases that can be categorized as aggregation visualizations.
- Years: The cases were published over a 19-year period from 1994 to 2013.
- Authors and datasets: Only Fernanda B. Viégas participated in more than one of the 10 cases in this category (cases 43 and 47); the rest participated in just one case each. The texts are very similar in nature to those in the item visualization category. Five cases are corpora that can be found online (*Wikipedia*, case 42; *Google*, cases 43 (figure 9) and 44; online news, case 46; search engine results, case 48). The standard unstructured texts include one from literature (*Sentence Bar Diagrams*, case 4), one from political speeches (*FeatureLens*, case 45) and one from a year's worth of email conversations between two correspondents (*Themail*, case 47). Finally, there are two quite unique cases: *Termite* (case 41) and *Dotplot* (case 49). All the cases are discussed below.

Discussion

Aggregation visualizations is the category with the greatest variation in the methods employed. Thus, apart from visualizing text collections, the only thing the 10 cases assigned to this category have in common is that they do not represent specific items.

Given these circumstances, we comment on each case separately:

Sentence bar diagrams (case 40) provide a matrix (or table-like) visualization that allows rows to be sorted by clicking on columns. The columns provide a quantitative definition of 13 metrics related to the 62 stories making up Grimm's fairy tales. It is a powerful tool for analysing, understanding and comparing the tales.

Termite (case 41) is a case that represents an intermediary dataset known as topic models. Topic models are a "cleverer" way of obtaining a bag-of-words from a text than applying a typical word-frequency statistical analysis. *Termite* does not visualize texts but it does compare parts of texts.

As such, the tool can be used to compare topic models.

Pediameter (case 42) is a specific interface that uses bar charts to show *Wikipedia* editions in real time. It is most remarkable for using a device known as an *Arduino* to detect editions and transcribe them to a physical indicator, merging digital and material worlds.

Web Seer (case 43) is another specific visualization method that shows the most popular search queries based on *Google* suggestions. The approach allows queries to be compared by representing the suggestions with trees and then connecting the matching branches. The simplicity of this case contrasts with its power of communication: rapid and user friendly.

Google's *tri-gram data* (case 44) uses a similar visualization method to that used by *Web seer*. It draws on the huge *Google* n-gram dataset and represents and compares three-word sentences (tri-grams).

FeatureLens (case 45) is an interactive, dashboard-style interface for comparing texts. The central representation uses a visualization of frequent concepts similar to that used by *Texty* (case 7) and *TileBars* (case 15). It allows text browsing and shows line graphs of frequent words found throughout a text.

Newsmap (case 46) uses treemap visualization to offer a new method for reading and monitoring the news in real-time, employing online *Google* news feeds. It is totally customizable in terms of topic, country and publication time. The software, which is available free of charge online, can also be used for news searches.

TheMail (case 47) is an experiment in which a highly specific interface was developed to follow and analyse the evolution of an email correspondence between two people over the course of one year. It visualizes the words that characterize each of the writers and their evolution over time.

When first developed in 1996, *WebBook* (case 48) (figure 10) was a somewhat surprising application, as it transformed search engine results in a multimedia (text and images, primarily) mash-up based on the metaphor of the book. The application was a pure text (web pages) collection visualization that presented the results as aggregations of text and images.

Finally, *Dotplot* (case 49) was an innovative visualization application with multiple uses, not unlike *Arc diagrams* (case 14). The main use of *Dotplots* is for text comparisons, including multi-language, text version and programming code comparisons.

c) Other subcategories

Here we include landscape data layers, search-result visualizations and time-dependent datasets.

Landscape as an additional data layer

Cases: 40, 26, 28, 33, 47, 37, 38 and 49.

The typical concept of landscape data is a network visualization comprising two layers of data, as in *Topic networks* (case 28). In this specific case, the first layer is provided by the Marcel Proust texts represented as items and the second layer by a network of topic models of these texts. The positions of the nodes of both layers are optimised so that proximity indicates more strongly related nodes. This definition of landscape can also be found in the defunct search engine results provided by *KartOO/Ujiko* (case 37) and *TouchGraph* (case 38).

Figure 10. (Case 48) *WebBook* by Stuart K. Card, George G. Robertson, and William York (1996)

All the other cases included in this category present text collections in combination with more data. This is the case of *Dotplot*, which represents the coincidence or otherwise of strings in various texts, and of *Grimm's fairy tale metrics*, which combines a list of texts in rows with various parameters listed in columns. These parameters do not form a direct part of the text, but rather they are recalculated features related to the text, including, for example, length, lexical diversity and the presence of different groups of words that represent entities (for example: body -> hand, head, heart, eyes and foot) in each tale.

> "Landscape visualizations allow to compare a collection of texts simultaneously with a second parameter"

A third kind of landscape is based on the representation of timed metadata, as exemplified by *Spot* (case 26), the *Royal Society Archive* (case 33) and *TheMail* (case 47).

A common feature of landscape visualizations is their capacity to compare a collection of texts simultaneously with a second parameter, while their main limitation is the number of items represented so that large numbers create problems of overlapping items.

Search result visualizations

Cases: 26, 43, 35, 45, 47, 46, 37, 38 and 48.

Compared to single-text visualizations, text-collection visualizations include considerably more cases offering search capacities (three vs. nine). Common sense suggests that when presenting a text collection, a natural feature of such an approach will be a way of selecting part of that collection based on given criteria, i.e., filter and search features.

All the cases included in this category allow search queries and output a unique visualization for each query. All the cases include a search box and a search button.

Time-dependent datasets

Cases: 42, 29, 36 and 46.

The four cases included in this category allow the user to monitor the evolution of the texts in the collection over time. Only one is designed for use in real-time (*Newsmap*, case 46), but potentially all of them can visualize the collection on a specific date and at a specific time.

One obstacle faced by an approach that represents changes in text collections over time is providing access to an updated feed or an accessible API. It is presumably for this reason that three of the four use *Wikipedia* data and the other uses *Google* news. In all cases, they are online sources that have long allowed public access to their feeds.

3. Conclusions

The diversity of approaches developed in different disciplines, the wide diffusion of publications or, on occasions, the absence of formal publications of innovative ideas, represent a considerable challenge to the undertaking of a comprehensive survey of the work completed in this field. Thus, some of the visualizations we present here have been unearthed in highly specific publications, the case for example of Joel Deshaye and Peter Stoicheff and their work on representing Faulkner (cases 11, 12 and 13). If we read Stoicheff's working notes it is apparent that their visualizations were developed to facilitate the study of William Faulkner's narrative timelines. There are no additional references to the application of these interesting ideas to other texts, suggesting that more works remain hidden in the depths of other fields.

Text visualization, as we have argued throughout this review, may be considered a subfield of data visualization. Yet, the boundaries of the discipline are not always clearly defined. This is readily illustrated, for example, by the case of Harrison's *Search clock* (case 31), in which the text corpora comprise an enormous dataset of search engine queries. Can this dataset really be considered a collection of texts when each of them, in most instances, is no more than one or two words in length? Does a text have to satisfy a minimum length in order to be considered a text? Here, we opted to treat case 31 as a collection of texts, short ones admittedly but, ultimately, *texts*.

Clearly, the critical decision to be made throughout this review has been how to classify the cases identified. As few papers have attempted to review only text visualization approaches, we turned to classic data visualization reviews (e.g., **Shneiderman**, 1996) as well as to more recent ones (e.g., **Collins** *et al*., 2009). In all these instances, the classifications were based on tasks that the visualization approach can solve rather than on the explicit aspects of the visualization themselves. For this reason we chose to propose our own classification, which, while far from perfect, we hope will be useful for undertaking a classification based on visual features.

We conclude with a list of insights, as well as shortcomings, that we have identified to date:

- Single-text visualizations have been applied mainly to literature, a field that, apart from being characterized by complex combinations of words, can present high levels of human abstraction and freedom of structure and experimentation. As such it might prove more effective to apply visualization techniques to texts that have a more formal register and/or predefined outline and a well-defined vocabulary, such as legal texts, scientific papers, template-based texts and communications, etc.
- We have identified only one single/partial-text visualization that is sequential (*Document arc diagrams*, case 22). Most partial-text visualizations extract the essence of the text based on one or more criteria and so the original sequence of the text is lost. Since sequential visualization approaches present certain advantages, it seems that partial-visualization approaches that maintain the original text sequence should be encouraged.
- Text-collection visualizations tend to employ methods that are used for data visualization in general. Hence, there is a need for further experimentation in applying more standard data visualization methods and approaches to the specific subfield of text visualization.
- Text collection aggregations is the category in which the most specific designs and ideas have been developed. More work needs to be undertaken to identify any common approaches in this kind of visualization.

And, finally, we pose the following question:

- Why is it that most of the cases reviewed here that are more than five years old are no longer available online? If the software used is no longer (or was never) in use, we should perhaps question its effectiveness. While we have not investigated just how many cases form part of commercial software products and how many, following publication, have simply been forgotten, the question remains as to why some apparently magnificent ideas did not establish themselves as new standards. Our challenge to researchers is to produce applications that will be adopted in one field or another, or which can solve a problem for a certain group of users; indeed, as the cases reviewed here highlight, adoption seems to represent a considerable challenge.

Acknowledgement

This work is part of the project "Active audiences and journalism. Interactivity, web integration and findability of journalistic information". CSO2012-39518-C04-02. *National plan for R+D+i, Spanish Ministry of Economy and Competitiveness*.

4. References

Anglin, Gary J.; **Vaez, Hossein**; **Cunningham, Kathryn L.** (2004). "Visual representations and learning: The role of static and animated graphics". *Handbook of research on educational communications and technology*, 2, pp. 865-916.

Baeza-Yates, Ricardo; **Ribeiro-Neto, Berthier** *et al.* (1999). *Modern information retrieval*. New York: ACM press, vol. 463.

Baeza-Yates, Ricardo; **Broder, Andreiz**; **Maarek, Yoelle** (2011). "The new frontier of web search technology: Seven challenges". *Search computing*, v. 6585 of *Lecture notes in computer science*, pp. 3-9.
http://dx.doi.org/10.1007/978-3-642-19668-3_1

Benavides, David; **Segura, Sergio**; **Ruiz-Cortés, Antonio** (2010). "Automated analysis of feature models 20 years later: A literature review". *Information systems*, v. 35, n. 6, pp. 615-636.
http://dx.doi.org/10.1016/j.is.2010.01.001

Collins, Christopher; **Carpendale, Sheelagh**; **Penn, Gerald** (2009). "DocuBurst: Visualizing document content using language Structure". *Computer graphics forum* (Procs. of the *Eurographics/IEEE-VGTC Symposium on visualization*, EuroVis), v. 28, n. 3, pp. 1039-1046.
http://dx.doi.org/10.1111/j.1467-8659.2009.01439.x

Feldman, Ronen; **Sanger, James** (2006). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. ISBN: 13 978 0 521 83657 9

Grobelnik, Marko; **Mladenić, Dunja** (2002). "Efficient visualization of large text corpora". In: *Procs of the 7th seminar*. Dubrovnik, Croatia.
http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf

Hearst, Marti A. (2003). *What is text mining?*
http://people.ischool.berkeley.edu/~hearst/text-mining.html

Hearst, Marti A. (2009). "Search user interfaces", Chapter 1. ISBN: 9780521113793
http://searchuserinterfaces.com/book
http://searchuserinterfaces.com/book/sui_ch1_design.html

Hearst, Marti A. (2011). "Natural search user interfaces". *Communications of the ACM*, v., 54, n. 11, November, pp. 60-67.
http://cacm.acm.org/magazines/2011/11/138216-natural-search-user-interfaces/fulltext
http://dx.doi.org/10.1145/2018396.2018414

Heer, Jeff (2010). "A conversation with Jeff Heer, Martin Wattenberg, and Fernanda Viégas". *Queue*, v. 8, n. 3, 10 pp., March.
http://doi.acm.org/10.1145/1737923.1744741

Iliinsky, Noah (2013). *Choosing visual properties for successful visualizations*. IBM Software. Business Analytics.
http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03323usen/YTW03323USEN.PDF

Kitchenham, Barbara (2004). *Procedures for performing systematic reviews*. Keele, UK, Keele University, 33 pp.

Levie, W. Howard; **Lentz, Richard** (1982). "Effects of text illustrations: A review of research". *ECTJ*, v. 30, n. 4, pp. 195–232.

Mann, Thomas M. (2002). *Visualization of search results from the world wide web*.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.2535

Meeks, Elijah (2011). *Digital humanities specialist*. Documents.
https://dhs.stanford.edu/comprehending-the-digital-humanities/documents

Nualart-Vilaplana, Jaume (2013). *How we draw texts: a visualization of text visualization tools.*
http://research.nualart.cat/textvistools

Nualart, Jaume; **Pérez-Montoro, Mario** (2013). "Texty, a visualization tool to aid selection of texts from search outputs". *Information research*, v. 18, n. 2, June.
http://www.informationr.net/ir/18-2/paper581.html

Shneiderman, Ben (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages*. Proceedings IEEE Symposium, pp. 336–343.
http://dx.doi.org/10.1109/VL.1996.545307

Šilić, Artur; **Dalbelo-Bašić, Bojana** (2010). "Visualization of text streams: A survey". *Knowledge-based and intelligent information and engineering systems*, v. 6277 of *Lecture notes in computer science*, pp. 31–43. Berlin, Heidelberg: Springer.
http://dx.doi.org/10.1007/978-3-642-15390-7_4

Stefaner, Moritz (2013). *Gender balance visualization*.
http://moritz.stefaner.eu/projects/gender-balance/#NUM/NUM

Strecker, Jacqueline (2012). *Data visualization in review: summary*. International Development Research Centre (IDRC), Ottawa, ON, Canada.
http://idl-bnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf

Times Higher Education. *World university rankings 2012-2013*.
http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking

Tufte, Edward R.; **Graves-Morris, P. R.** (1983). *The visual display of quantitative information*, v. 2. Cheshire, CT: Graphics Press, 199 pp.